A robust subspace classification scheme based on empirical intersection removal and sparse approximation

Integrated Computer-Aided Engineering 22 (2015) 59–69
DOI 10.3233/ICA-140470
IOS Press

A robust subspace classification scheme based on empirical intersection removal and sparse approximation

Yinan Yu* and Tomas McKelvey
Chalmers University of Technology, Gothenburg, Sweden

Abstract. Subspace models are widely used in many applications. By assuming an individual subspace model for each class, linear regression is applied and combined with minimum distance criteria for making the final decision. In a generalized subspace model, the full linear subspace of each class is split into subspaces with lower dimensions, and the unknown basis needs to be estimated with respect to the testing pattern using adaptively selected training samples. The training data selection is implemented using either least-squares regression or sparse approximation. In this paper, to further improve the classification performance, instead of attempting to minimize the regression error for each class, the between class separability is enhanced by a novel approach called Empirical Subspace Intersection (ESI) Removal technique. Evaluations are performed on (1) standard UCI data set, and (2) a computer aided system along with the proposed classification technique to determine the quality in wooden logs using microwave signals. The experimental results are shown and compared with classical methods.

Keywords: Classification, linear subspace, sparse representation, training data selection

1. Introduction

Assuming that the data generation mechanism can be described using subspace models, the classification problem is hence the identification of the subspace to which that the testing patterns belong. When the dimensionality of the feature space is much larger compared to the number of available training samples, the global topological properties and statistical assumptions become extremely difficult to verify and the training phase of the classification algorithm becomes very challenging. This issue is referred to as a "High Dimensionality and Low Sample Size" (HDLSS) [27] problem, where linear subspace models are usually chosen to avoid overfitting and enhance robustness.

In this paper, a classifier based on linear subspace settings has been developed, with the assumption that a pattern from one class lies on *one of several possible linear subspaces*. Hence, we regard each class to be associated with a set of subspaces. We call it the "generalized subspace model" [37], which explores the data structure in a localized fashion. More precisely, the subspaces are estimated in the following way. Given a testing pattern, there are two steps involved:

(1) The active training data are identified using a least-squares criterion or a sparse approximation;
(2) The selected training data are used to estimate the span of the sub-basis.

These procedures are carried out individually for all classes and the decision is based on minimum distance or maximum projection length. To further improve the robustness, the class separability is enhanced by removing the Empirical Subspace Intersection (ESI). ESI is a subspace with a given dimensionality, which has the smallest principal angles to the subpace of the other class.

The proposed classification approaches are applied to a wood quality assessment problem based on sig-

*Corresponding author: Yinan Yu, Chalmers University of Technology, 412 96 Gothenburg, Sweden. E-mail: yinan@chalmers.se.

ISSN 1069-2509/15/$35.00

nals from a microwave array sensor. Microwave signals are widely used for applications in different areas [10,15,18,25,28,34], based on which, computer aided systems are developed for assisting in human decisions. Such experimental signals result in a typical HDLSS data set, and it is empirically verified that our proposed scheme provides a robust solution.

This paper is organized as follows. A short review on the subspace model and existing techniques is given in Section 2.1. The generalized subspace model is introduced in Section 2.2 and the corresponding classification scheme is proposed in Section 3. Application and empirical results are presented in Section 4, where comparison with detailed parameter descriptions are given and discussed.

2. Signal model and classification hypothesis

Given C the total number of classes, let $\{\boldsymbol{x}_c^i\}$ be the training set of class $c \in \{1,\ldots,C\}$, where $\boldsymbol{x}_c^i$ is p dimensional complex valued, $i \in \{1,\ldots,N_c\}$ is the sample index and N_c is the number of training samples from class c. In this paper, let us focus on binary classification problems, i.e. $c \in \{1,2\}$.

2.1. Review of linear subspace model

In a linear subspace model, each data point $\boldsymbol{x}_c$ drawn from class c is assumed to be generated from the following generating function:

$$\boldsymbol{x}_c = \boldsymbol{U}_c\boldsymbol{\alpha}_c + \boldsymbol{e} \tag{1}$$

where the columns of $\boldsymbol{U}_c$, denoted as $\{\boldsymbol{u}_{c,l}\}$ represent the orthonormal basis of the corresponding linear subspace with $l \in \{1,\ldots,D_c\}$; $\boldsymbol{\alpha}_c$ is the weighting vector; and $\boldsymbol{e}$ is zero mean random noise.

With a given $\boldsymbol{U}_c$, we can compute the distance $d_c(\boldsymbol{x})$ from any signal vector $\boldsymbol{x}$ to the linear subspace spanned by its orthonormal columns $\{\boldsymbol{u}_{c,l}\}$:

$$d_c(\boldsymbol{x}) = \|\boldsymbol{x}-\boldsymbol{P}_c\boldsymbol{x}\|_2+\gamma_c = \left\|\boldsymbol{x}-\boldsymbol{U}_c\boldsymbol{U}_c^H\boldsymbol{x}\right\|_2+\gamma_c \tag{2}$$

where $\boldsymbol{P}_c$ denotes the projection matrix and $\boldsymbol{U}_c^H$ is the Hermitian transpose of the matrix $\boldsymbol{U}_c$. Here γ_c is a biasing parameter which, for the binary classification problem, can be used to control the probability of detection and false alarm rate tradeoff.

If signal $\boldsymbol{x}$ is unlabeled, the class label $\hat{c}$ can be estimated according to the following criterion:

$$\hat{c} = \arg\min_c d_c(\boldsymbol{x}) \tag{3}$$

In practice when the basis $\boldsymbol{U}_c$ is unknown, the classifier associated with the model assumption given in Eq. (1) and criterion in Eq. (3) can be constructed by defining the projection matrix $\boldsymbol{P}$ as:

$$\boldsymbol{P}_c = \boldsymbol{X}_c(\boldsymbol{X}_c^H\boldsymbol{X}_c)^{-1}\boldsymbol{X}_c^H \tag{4}$$

where, data matrix $\boldsymbol{X}_c$ is constructed by placing all training data from class c as its columns:

$$\boldsymbol{X}_c = \left[\boldsymbol{x}_c^1,\ \boldsymbol{x}_c^2,\ldots,\boldsymbol{x}_c^{N_c}\right] \tag{5}$$

In the literature, under the model assumption introduced in Eq. (1), a signal detection technique called Matched Subspace Detector (MSD) [4,19,29] has been proposed and analyzed based on the derivation of Generalized Likelihood Ratio Test (GLRT) [38]. Different scenarios are taken into consideration, such as known/unknown basis $\boldsymbol{U}_c$, covariance matrix, etc. In the machine learning community, the corresponding classification method is the so-called CLAss-Featuring Information Compression (CLAFIC) [24]. Its nonlinear counterpart is presented in [2]. Another branch of data driven classification techniques named Linear Regression Classifier (LRC) [21] have been presented for face recognition in 2010, which essentially applies the same criteria as in [24]. Modifications and extensions of LRC are developed accordingly, such as Principal Component Regression Classifier (PCRC), Improved Principal Component Regression Classifier (IPCRC) [16], Robust LRC [22], Ridge Regression Classifier (RRC) [1], Unitary Regression Classifiers (URC) [17], etc. These existing techniques generalize the idea of LRC. For example, PCRC and IPCRC tend to improve the performance by manipulating the principal components in the PCA space; RRC is developed to handle degenerated cases using ridge regression and URC tries to minimize the total within class projection error.

These techniques are all derived from the model in Eq. (1), which does not necessarily hold in reality. In some applications, more than one subspace is involved in the data generating process. However, even though generated from different subspaces, some signals might be grouped together into one "super" class when we create the training data. For example, when

we try to distinguish "animals" from "cars" in images, the "animal" could be a cat, a dog or an elephant, and we assume that each kind of animal corresponds to one subspace. In this case, we call the class "animal" a "super class". Therefore, each "super class" contains the union of several subspace models which need to be estimated. Below we introduce a generalized subspace model which will account for classes where data are described by a union of subspaces.

2.2. Generalized subspace model

In a generalized subspace model, instead of a linear subspace spanned by $\boldsymbol{U}_c$, each $\boldsymbol{x}_c$ is considered to be generated from one out of a set of linear subspaces spanned by the 'smaller' basis $\boldsymbol{U}_c^k$, where $k \in \{1, \ldots, K_c\}$, and K_c is the total number of such subspaces. By 'smaller' basis, one can imagine that the subspace spanned by the basis appeared in Eq. (1) is now a set of K_c linear subspaces spanned by some low dimensional bases.

Definition 1 (Generalized subspace model (GSM)). Let $\mathscr{U}_c = \{\boldsymbol{x} : \boldsymbol{x} \in \text{class } c\}$. We assume:

$$\mathscr{U}_c = \bigcup_{k \in \{1 \ldots K_c\}} \mathscr{U}_c^k \tag{6}$$

with

$$\mathscr{U}_c^k = \left\{ \boldsymbol{x} : \ \boldsymbol{x} = \sum_{l=1}^{D_c^k} \beta_l \boldsymbol{u}_{c,l}^k \right\}, \ \ \beta_l \in \mathbb{C} \tag{7}$$

where D_c^k is the dimension of the subspace $\boldsymbol{U}_c^k = \left\{ \boldsymbol{u}_{c,l}^k \right\}$ with $l \in \{1, \ldots, D_c^k\}$, and β_l is the corresponding coefficient. □

Hence, each class is defined as a union of K_c subspaces where each subspace is represented by the unitary matrix $\boldsymbol{U}_c^k$. The signal model in Eq. (1) thus becomes:

$$\boldsymbol{x}_c = \boldsymbol{U}_c^{k_c} \boldsymbol{\beta}_c + \boldsymbol{e} \tag{8}$$

By denoting $d_c^k(\boldsymbol{x})$ the distance from $\boldsymbol{x}$ to the subspace $\boldsymbol{U}_c^k$, we have:

$$d_c^k(\boldsymbol{x}) = \left\| \boldsymbol{x} - \boldsymbol{U}_c^k (\boldsymbol{U}_c^k)^H \boldsymbol{x} \right\|_2 + \gamma_c \tag{9}$$

and $\hat{c}$ can be estimated in the same way as in Eq. (3) with a slight modification.

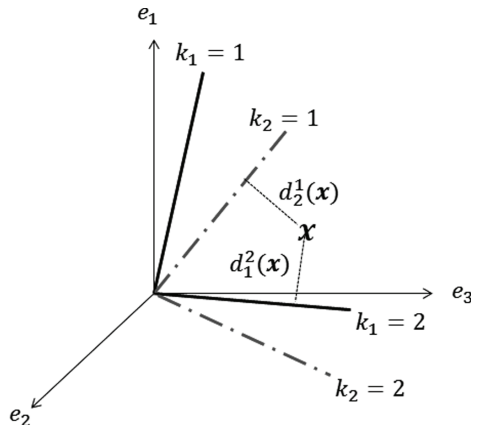


Fig. 1. A three dimensional example is shown. The solid and dash–dotted lines indicate the linear subspaces from class 1 and 2 respectively (of dimension 1) closest to the data point $\boldsymbol{x}$.

$$\hat{c} = \arg \min_c \min_k \ d_c^k(\boldsymbol{x}) \tag{10}$$

A low dimensional example is visualized in Fig. 1. Data point $\boldsymbol{x}$ represents a high dimensional vector (here three dimensional for convenience). The solid and dash-dotted lines represent the two one-dimensional subspaces contained in class 1 and 2 respectively. In the figure, subscript c of k_c indicates the class number. The symbol $d_c^k(\boldsymbol{x})$ indicates the distance from $\boldsymbol{x}$ to the corresponding subspace $\boldsymbol{U}_c^k$.

The full subspace basis representations for each class, $\boldsymbol{U}_c^k$, are normally *not known* and have to estimated from the labeled training data.

3. Proposed method

The basic idea behind the proposed algorithm is that, given a testing sample $\boldsymbol{x}$, we try to establish which of the training vectors from each class should be included in the training process. The classification algorithm can be divided into 4 phases:

1) Adaptive training data selection.
2) Subspace basis estimation.
3) Empirical Intersection removal.
4) Classification based on minimum distance.

The classification scheme including steps 1), 2) and 4) is called Generalized Subspace Regression Classifier (GSRC) and along with step 3), the method is called Improved GSRC (IGSRC). The flow chart of the proposed classification scheme is shown in Fig. 2 and each step is explained and discussed in detail below.

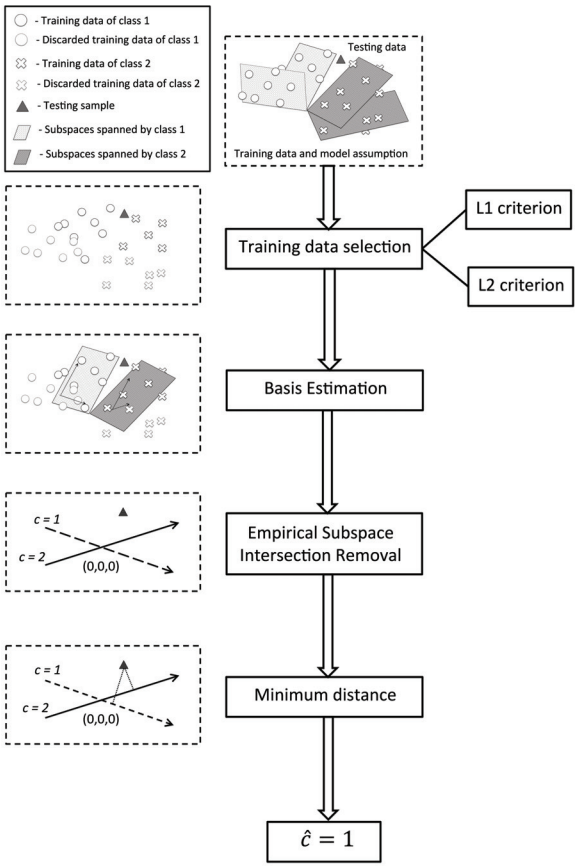


Fig. 2. The flow chart of improved nearest subspace regression classifier.

3.1. Step 1: Adaptive training data selection for $\boldsymbol{x}$

According to Eq. (8), each sample from class c is assumed to be lying in one of the K_c subspaces. Therefore, before computing the distance by Eq. (9), we need to select the 'correct' training set for $\boldsymbol{x}$ in both classes. Assuming the subspace dimension D_c is known and fixed, we can formulate the optimal selection as the D_c vectors from all training data which give the smallest projected distance. This can be illustrated in Fig. 1. As we can see, given an unlabeled data $\boldsymbol{x}$, although the lines with the same type represent the same class, we still need to select one of them to compute the distance for $\boldsymbol{x}$. In this case, the selected subspaces are lines $k_2 = 1$ and $k_1 = 2$.

Given data matrix $\boldsymbol{X}_c$ defined in Eq. 5, the task is to select D_c relevant columns from $\boldsymbol{X}_c$, such that $\boldsymbol{x}$ can be approximated by a linear combination of the basis for the subspace spanned by these data. The selection is carried out by computing a weighting vector $\boldsymbol{w}_c$ whose j^{th} element represents the importance of column j in $\boldsymbol{X}_c$, (i.e. $\boldsymbol{x}_c^j$) with respect to reconstructing the $\boldsymbol{x}$. The most significant vectors are then chosen to be the correct training data of $\boldsymbol{x}$. Below we present two computational methods to perform the selection.

– *Formulation using least square criterion* Here we simply derive the least-squares solution,

$$\boldsymbol{w}_{\text{LS}} = \arg\min_{\boldsymbol{w}_c} \|\boldsymbol{x} - \boldsymbol{X}_c\boldsymbol{w}_c\|_2^2 \tag{11}$$

and select the training data as the vectors in $\boldsymbol{X}_c$ corresponding to the indices of the D_c elements of $\boldsymbol{w}_{\text{LS}}$ with the largest magnitude. Although computationally simple, it is well known that this method yields a suboptimal solution to the regressor selection problem since all elements in $\boldsymbol{w}_{\text{LS}}$ in general is non-zero.

– *Formulation using a sparse approximation* To obtain a sparse solution, the approximation criterion need to include a term which promote a sparse solution [5,11,14,40]. If the least-squares criterion is augmented with an l_1 penalty on the weight vector

$$\boldsymbol{w}_{\text{L1}} = \arg\min_{\boldsymbol{w}_c} \|\boldsymbol{x} - \boldsymbol{X}_c\boldsymbol{w}_c\|_2^2 + \lambda\|\boldsymbol{w}_c\|_1 \tag{12}$$

the solution will be sparse, and the parameter λ controls the level of sparseness. Equation (12) is the Lagrangian formulation of the "least absolute shrinkage and selection operator" (LASSO) method [31] and can be solved, for a given λ, with convex optimization algorithms [6]. By performing a line search over λ, a solution with desired level of sparseness, i.e. D_c nonzero components in $\boldsymbol{w}_{\text{L1}}$, can be obtained.

Figures 3 and 4 illustrate an example of the magnitude of the weight vector resulting from the least-squares and sparse approach respectively. It is evident that the sparse approach yields many zero components in the weight vector. The vectors in $\boldsymbol{X}_c$ corresponding to the significant elements of $\boldsymbol{w}$ are selected to construct the matrix with columns spanning the subspace $\boldsymbol{O}_c$. Let J denote the set of the indices of non-zero elements of $\boldsymbol{w}_{l1}$, or the D_c elements of $\boldsymbol{w}_{LS}$ with largest magnitude. Then

$$\boldsymbol{O}_c = \left[\boldsymbol{x}_c^j\right]_{j\in J}. \tag{13}$$

This sparsity learning model is closely related to the model presented in Sparse Representation-based Clas-

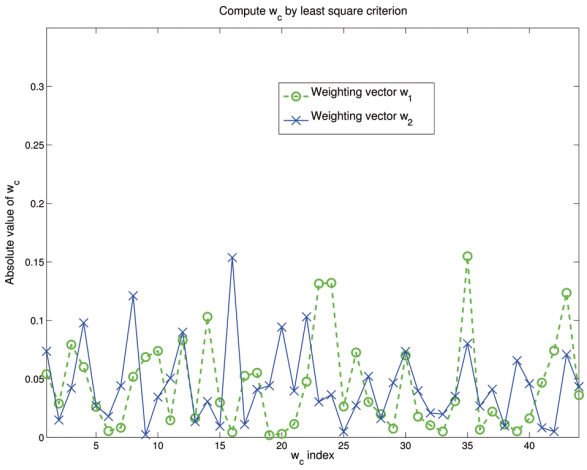


Fig. 3. The weighting vector $\boldsymbol{w}_c$ estimated with respect to least squares criterion.

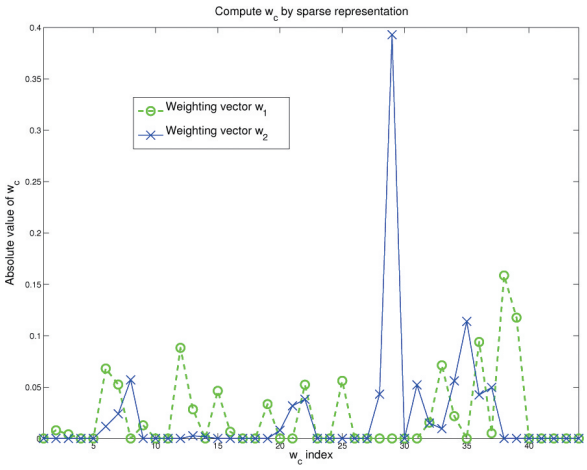


Fig. 4. The weighting vector $\boldsymbol{w}_c$ estimated using sparse approximation.

sification (SRC) technique [35]. However, in [35] the test pattern is sparsely reconstructed using training patterns from all classes simultaneously, i.e. no individual sparse subspace for each class is assumed. Improvements are presented in [9], where a block based reconstruction is developed with a better representation of the testing pattern. On the other hand, [20,33] and the corresponding kernel techniques [7,39] provide similar, but different, selection schemes based on nearest neighbor criterion in the subspace sense.

3.2. Step 2: Basis estimation for $\boldsymbol{O}_c$

Let $\boldsymbol{O}_c$ be the matrix constructed by Eq. (13). An orthonormal basis of the corresponding subspace can be estimated by the left singular vectors of $\boldsymbol{O}_c$ computed from the singular value decomposition (SVD). The first steps in Algorithm 1 show the details.

3.3. Step 3: Empirical subspace intersection removal

After obtaining the basis vectors determining the subspace for each class, a classification criterion can be applied. However, empirical testing results suggest that by removing the directions with small principal angles [12] between the two selected basis, the classification performance can be improved. To fully understand this concept, first we define the principal angles as follows:

Definition 2. *Principal angles*

The principal angles θ_k between the subspaces $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are defined as:

$$\begin{aligned} \cos(\theta_k) = \max_{\boldsymbol{u}\in\boldsymbol{U}_1, \boldsymbol{v}\in\boldsymbol{U}_2} & \boldsymbol{u}^H\boldsymbol{v} = \boldsymbol{u}_k^H\boldsymbol{v}_k \\ \text{subject to:} & \\ & \|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1 \\ & \boldsymbol{u}^H\boldsymbol{u}_i = 0, \quad i = 1,\ldots,k-1 \\ & \boldsymbol{v}^H\boldsymbol{v}_i = 0, \quad i = 1,\ldots,k-1 \end{aligned} \tag{14}$$

where $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$ are the principal vectors corresponding to the $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ spaces respectively

The principal angle is a distance metric. More precisely, the distance between two subspaces is considered small when the principal angle (s) are small. Particularly, if p principal angles are zero then the two spaces $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ share a p dimensional common subspace, *the intersection*. Since each subspace represents an individual class, this distance thus reflects the class separability to some extent. We define the space spanned by the principal vectors of a subspace corresponding to the smallest principal angles as the Empirical Subspace Intersection (ESI) between the two spaces. It is called empirical due to the fact that we identify these components from estimated subspaces which typically do not have non-zero intersection. ESI is defined as follows:

Definition 3. Empirical Subspace Intersection

The Empirical Subspace Intersection $ESI(\boldsymbol{U}_1, \boldsymbol{U}_2, \delta)$ of $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ is defined as:

$$\begin{aligned} & ESI(\boldsymbol{U}_1, \boldsymbol{U}_2, \delta) = \\ & \quad \{\boldsymbol{u}_j : \boldsymbol{u}_j \in \boldsymbol{U_1}, \ \forall\, j,\ s.t.\ \theta_j < \delta\} \end{aligned} \tag{15}$$

where s θ_j's are the principal angles and $\boldsymbol{u}_j$ are the principal vectors and δ is the empirical tolerance. □

Note that ESI is not symmetrically defined for both subspaces, since $ESI(\boldsymbol{U}_i, \boldsymbol{U}_j, \delta) \subset \boldsymbol{U}_i$ holds, but not necessarily $ESI(\boldsymbol{U}_i, \boldsymbol{U}_j, \delta) \subset \boldsymbol{U}_j$. Thus the order of the subspaces in the bracket reflects which subset we are selecting from.

The algorithm of removing ESI from pre-computed bases $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ is summarized in the second part of Algorithm 1 (see also [12]).

After this step, the basis $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are obtained and minimum distance classification criterion can be applied according to Eqs (9) and (10).

3.4. Algorithm

The algorithm is presented in Algorithm IGSRC.

In summary, as opposed to the previous techniques, the classification technique proposed in this paper is different in the following ways: 1) In [35] a single sparse regression is performed using training-data from all classes and in a second step do the classification based on the sparse representation vector. Here we perform a sparse regression for each class. 2) As suggested by the name, LRC and extensions emphasize on the 'goodness of fit' in a regression sense. However, the main focus of our method is to enhance the class separability by removing the subspaces in the basis corresponding to small principal angles.

4. Applications and results

The presented method is evaluated using data from two different classification applications; wood quality assessment and cancer detection from mass-spectroscopy data.

4.1. Application on wood quality assessment

4.1.1. Signal description

The classification problem studied in this example is to detect rot in wooden logs using data from a microwave measurement system. An illustration of the setup is shown in Fig. 5. Each green cross indicates the position of one antenna which can act both as a transmitter and a receiver. The transmitter and the receiver number are indicated by q and p respectively, and the pair is referred as channel $\{p, q\}$. The scattering parameter, for a given excitation frequency, at each channel (p, q) is the denoted by S_{pq}, and is defined as:

$$S_{pq} = -\frac{H^-_{0,p}}{H^+_{0,q}} \tag{22}$$

Algorithm 1 ESI removal between two subspaces defined by the matrices $\boldsymbol{O}_1$ and $\boldsymbol{O}_2$:

Note: in this algorithm, the intermediate basis is called $\boldsymbol{Q}_c$, $c = 1, 2$. The notation $\boldsymbol{U}_c$, $c = 1, 2$ is used to denote the final constructed basis for class c.

- Let $\boldsymbol{O}_1$ and $\boldsymbol{O}_2$ be of sizes $p \times M_1$ and $p \times M_2$ respectively. Without loss of generality we assume $D_1 \geqslant D_2$;
- Compute the SVD factorization:
$$\boldsymbol{O}_c = \boldsymbol{Q}_c \boldsymbol{S}_c \boldsymbol{V}_c^H$$
- Truncate $\boldsymbol{Q}_c$ to the predefined dimension D_c if $M_c > D_c$:
$$\boldsymbol{Q}_c \leftarrow \boldsymbol{Q}_c(:, 1 : D_c) \tag{16}$$
- Construct matrix $\boldsymbol{C}$:
$$\boldsymbol{C} = \boldsymbol{Q}_1^H \boldsymbol{Q}_2 \tag{17}$$
- Compute the full SVD of $\boldsymbol{C}$:
$$\boldsymbol{Y}^H \boldsymbol{C} \boldsymbol{Z} = \begin{bmatrix} \cos(\theta_1) & \dots & 0 \\ 0 & \ddots & 0 \\ 0 & \dots & \cos(\theta_{D_2}) \\ 0 & 0 & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0 & 0 \end{bmatrix} \tag{18}$$
where θ_k's are the principal angles and $1 \geqslant \cos(\theta_1) \geqslant \cdots \geqslant \cos(\theta_s) \geqslant \cos(\theta_{s+1}) \geqslant \cdots \geqslant \cos(\theta_{D_2}) \geqslant 0$, where s denotes the dimension of the subspace to be removed.
- Compute the associated basis $\{\boldsymbol{u}_k\}$ and $\{\boldsymbol{v}_k\}$:
$$\boldsymbol{Q}_1 \boldsymbol{Y} = [\boldsymbol{u}_k], k = 1, \dots, D_1$$
$$\boldsymbol{Q}_2 \boldsymbol{Z} = [\boldsymbol{v}_k], k = 1, \dots, D_2 \tag{19}$$
and
$$ESI(\boldsymbol{Q}_1, \boldsymbol{Q}_2, \theta_{s+1}) = \left[\boldsymbol{u}_1, \boldsymbol{u}_2, \dots \boldsymbol{u}_s\right]$$
$$ESI(\boldsymbol{Q}_2, \boldsymbol{Q}_1, \theta_{s+1}) = \left[\boldsymbol{v}_1, \boldsymbol{v}_2, \dots \boldsymbol{v}_s\right] \tag{20}$$
- Remove the intersection and construct the subspaces:
$$\boldsymbol{U}_1 = \left[\boldsymbol{u}_{s+1}, \boldsymbol{u}_{s+2}, \dots, \boldsymbol{u}_{D_1}\right]$$
$$\boldsymbol{U}_2 = \left[\boldsymbol{v}_{s+1}, \boldsymbol{v}_{s+2}, \dots, \boldsymbol{v}_{D_2}\right] \tag{21}$$

Algorithm IGSRC

- Produce the feature vector $\boldsymbol{x}$ by pre-processing the data;
- $\forall c \in \{1, 2\}$, compute the weight vector $\boldsymbol{w}_c$:

$$\boldsymbol{w}_c = \arg \min_{\boldsymbol{w}_c} \|\boldsymbol{x} - \boldsymbol{X}_c \boldsymbol{w}_c\|_2^2 + \lambda \|\boldsymbol{w}_c\|_1$$

  where λ should be obtained by line search as mentioned before.
- Select the D_c elements of $\boldsymbol{w}_c$ with the largest magnitude and identify the set of corresponding indices J_c of $\boldsymbol{X}_c$;
- Construct the matrix:

$$\boldsymbol{O}_c \leftarrow \left[\boldsymbol{x}_c^j\right], \quad j \in J_c$$

- Basis estimation and ESI removal according to Algorithm 1 to derive $\boldsymbol{U}_c$
- Compute the distance:

$$d_c(\boldsymbol{x}) = \left\| \boldsymbol{x} - \boldsymbol{U}_c \left(\left(\boldsymbol{U}_c\right)^H \boldsymbol{x} \right) \right\|_2$$

- Estimate the label of $\boldsymbol{x}$:

$$\hat{c} = \arg \min_c d_c(\boldsymbol{x})$$

Fig. 5. An illustration of the experimental setup. Each cross indicates the position of one antenna, playing a role both as transmitter and receiver. The signal is then measured as the S parameters in frequency domain. The antennas are labeled as $1, 2, \ldots, N_a$ in a counterclockwise order.

where $H_{0,p}^-$ is the complex amplitude of the fundamental mode of the outgoing wave at port p and $H_{0,q}^+$ is the system excitation amplitude at port q.

The S_{pq} measured at frequency f_n [Hz] can be written in the following form:

$$S_{pq}(f_n) = e^{\eta + j\phi} \tag{23}$$

where the real part η represents the amplitude and the imaginary part ϕ represents the phase information. Since we assume the system is reciprocal, $S_{pq}(f_n) = S_{qp}(f_n)$ holds.

By considering all the frequency points measured, the signal for a given channel $\{p, q\}$ can be expressed as the row vector:

$$\boldsymbol{S}_{pq} = \left[S_{pq}(f_1), \ldots, S_{pq}(f_{N_f})\right] \tag{24}$$

4.1.2. Experimental results

Combing all channels and frequencies we obtain the data vector:

$$\boldsymbol{x} = \left[\boldsymbol{S}_{11}, \ldots, \boldsymbol{S}_{1\ N_q}, \boldsymbol{S}_{22}, \ldots, \boldsymbol{S}_{2\ N_q}, \ldots \boldsymbol{S}_{N_p N_q}\right]^T \tag{25}$$

Furthermore, according to our setup, we have $N_p = N_q = N_a$, where N_a is the number of antennas. The vectorized signal $\boldsymbol{x}$ is therefore considered as a p dimensional vector, where the dimension p is given by

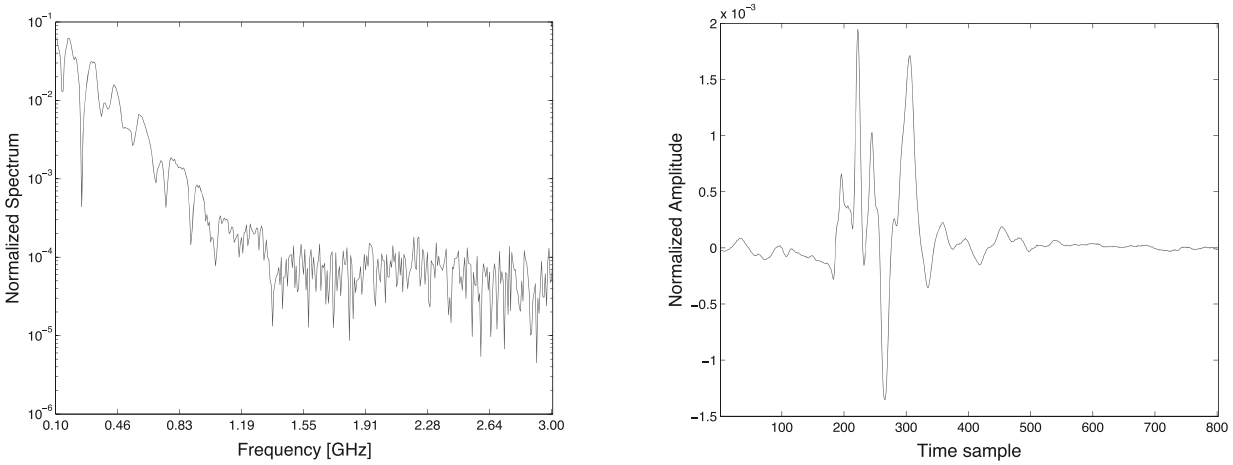
$$p = \frac{1}{2} N_f (N_a^2 + N_a). \tag{26}$$

In this work, $N_f = 180$ frequency points (corresponding to approximately $0.1 \sim 1.3$ GHz) are used for classification. The number of antennas is $N_a = 12$. One example of the measurements can be found in Fig. 6. From Eq. (26), we have the dimension of the vectorized signal $\boldsymbol{x}$ is 14040. In this experiment, 54 and 108 samples for normal and rotten timbers are measured respectively.

4.1.3. Pre-processing

Different types of signal pre-processing procedures can be applied before the signals are used as the input of the classifier. In this work, the main operations are 1) logarithm transformation; 2) normalization of each channel.

- Logarithm transform
  We take $\log(\boldsymbol{x})$ as the new signal vector instead of the $\boldsymbol{x}$ defined in Eq. (25) to retrieve the complex number $\eta + j\phi$.
- Normalization
  The reflection S_{pq}, where $p = q$, is typically much stronger than the transmission where $p \neq q$. However, the later one might carry more information of the object. Therefore, to unify the contribution of different channels, a channel-wise normalization is implemented on the signal $\boldsymbol{S}_{pq}$ in Eq. (24)



(a) S parameter measured at channel $\{1, 5\}$ (b) Corresponding time domain signal

Fig. 6. Left graph (a) The absolute value of the measured S parameters $\boldsymbol{S}_{1,5}$ using the antennas at the 1^{st} and the 5^{th} positions as input and output sensors respectively over all the frequency points. Right graph (b) The corresponding time domain signal.

to ensure that they contain the same energy level.

$$\boldsymbol{S}'_{pq} = \frac{\log \boldsymbol{S}_{pq}}{\|\log \boldsymbol{S}_{pq}\|_2} \tag{27}$$

4.1.4. Results and discussion

A randomized testing procedure has been carried out, where the samples are randomized, among which 44 are used for training and the rest for testing. The randomization is repeated 30 times. The results are evaluated by the rotten wood detection rate at a constant false alarm of 25%, and compared with classical methods such as Support Vector Machines (SVM) [8, 32], K Nearest Neighbors (KNN) [3], Linear Discriminant Analysis (LDA) [26] and existing subspace classifiers. To compare these approaches, the detection rate and their parameters are shown in Table 1. The effect of different formulations of $\boldsymbol{w}_c$, discussed in Section 3.1, on the classification result (referred as LS and L1 in Table 1) is similarly compared.

As we can see from the evaluation, KNN classification scheme does not provide a good detection result. SVM based classifiers are not very efficient in our experiment. When the dimensionality of the feature space is extremely large compared to the sample size available, the performance of SVMs might be degraded [23]. Nevertheless, by applying a RBF kernel to SVM [30], the performance improves compared to the linear case. Finally, the results of linear discriminant analysis in the PCA space does not show any improvement compared to linear SVM.

Among the subspace classifiers, IPCRC does not perform better than PCRC in our experiment. Since IPCRC discards the subspace spanned by the first principal components, the result might be degraded when these subspaces contain important discriminant information. On the other hand, by discarding the directions corresponding to small eigenvalues in the principal component space, PCRC considers eigenvectors corresponding to small eigenvalues as random directions and information may be lost as well. Instead, GSRC improves the results by assuming generalized subspace model and selects training data under L1 or LS constraints. By removing empirical subspace intersection between the subspaces, the performance is further enhanced (IGSRC).

4.1.5. Settings of parameters

The parameters in this experiment are chosen by cross-validation using 10% of the training data. Precisely speaking, leave-one-out validation technique is applied to choose the parameters by maximizing the performance on this 10% and the rest of the data set is used for training and testing.

Note that the parameter D_c is used in multiple algorithms. We therefore tuned D_c with respect to the best performance of the classical method PCRC to have a fair comparison.

In this section, we discuss the effects of the parameters to the classification performance.

- Dimension of ESI: s
  To further study how the dimensionality of ESI affects the results, the detection rate at a 25% constant false alarm rate is shown in Fig. 7. As we can see from the empirical results, there is an opti-

Table 1
The detection rate of rotten log at a constant false alarm rate (25%) obtained by randomized N-testing (c.f. 4.1.3)

	Comparison of results		
Method	Parameters	Parameter value	Detection rate
KNN	K	10	64.1%
linear SVM	C	10	75.5 %
RBF kernel SVM	C, σ	10, 0.5	77.2%
PCA + LDA	Number of principal components	25	73.4%
LRC	–	–	78.6%
PCRC	Number of principal components	25	79.2%
IPCRC	Number of discarded (first) principal components	10	78.5%
GSRC (LS)	Subspace dimensionality D_c	25	81.4%
GSRC (L1)	Subspace dimensionality D_c	25	85.6%
IGSRC (L1)	Subspace dimensionality D_c	25	89.1%
	(removed) ESI dimensionality	6	

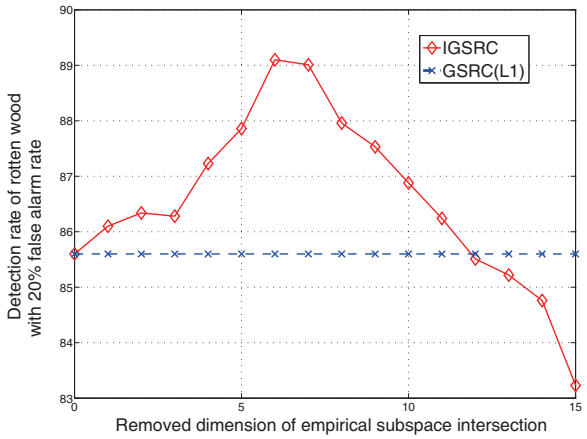


Fig. 7. The effects of the ESI dimension on the classification performance (at a constant false alarm rate of 25%).

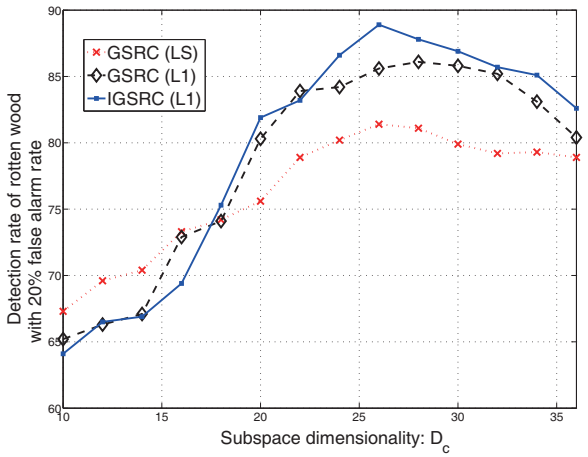


Fig. 8. The effects of the basis dimension D_c on the detection performance (at a constant false alarm rate of 25%).

mum dimension for the removed ESI. The reason is that when dim(ESI) is high, it implies a large tolerance value δ from Eq. (15). The larger δ is, the more probable that the removed subspace contains discriminant information between the two subspaces. Therefore, by removing ESI with a too large δ, the performance can be degraded.

– Dimension of subspace $\boldsymbol{U}_c$: D_c Another important parameter is the dimension of the subspace $\boldsymbol{U}_c$ for class c. Note that here we have $D_1 = D_2 = D_c$, but in principle they do not have to be chosen the same. In Fig. 8, the performances of the classifiers GSRC (LS & L1) and IGSRC (L1) are plotted as a function of the predefined subspace dimension D_c. From the figure, we observe that:

  * $D_c < 18$: IGSRC does not perform as well as GSRC (L1). ESI is considered containing redundant directions from a classification point of view. Therefore, with a very small dimensional subspace, this redundancy is very difficult to identify and hence information may be lost by removing any subspace form the two classes.
  * $D_c = 26$: IGSRC reaches the best performance. When the dimension D_c keeps growing, the advantage of training data selection is degraded. When $D_c = N_c$, the whole data set is used for training and there will be no selection at all.
  * $D_c < 20$: no advantages can be concluded by using L1 constraints instead of LS for training data selection. However, by allowing higher D_c, the performance of GSRC (L1) exceeds

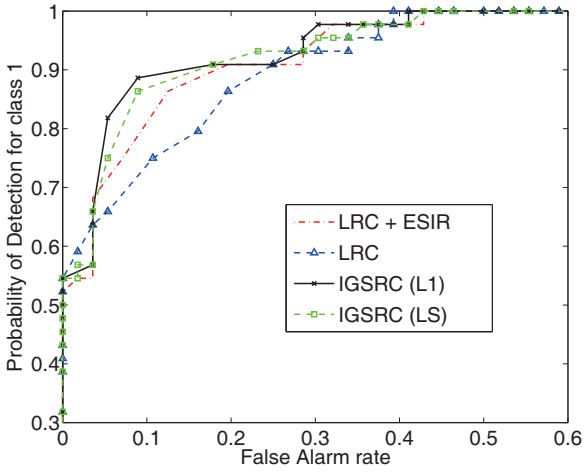


Fig. 9. The ROC curve achieved using different approaches on UCI data set ARCENE.

GSRC (LS), since L1 representation is more advanced for selecting relevant directions.

4.2. Test on UCI data set ARCENE

The algorithm is also applied to a standard UCI data set "ARCENE" [13], where $\boldsymbol{x} \in \mathbb{R}^{10000}$. Here the problem is to detect cancer based on mass-spectroscopy measurements. This is a typical HDLSS data set where only 100 patterns are available for training. The result shown is based on the testing error of a validation data set with 100 samples. Results in terms of False Alarm rate versus Probability Detection is shown in Fig. 9. As we can see that compared to traditional LRC, Empirical Subspace Intersection Removal (ESIR) technique provides a better result. Moreover, the performance is further improved by assuming Generalized Subspace model, especially with sparse approximation for training data selection.

5. Conclusion

In this paper, a classification scheme has been proposed, which includes four steps: 1) adaptive training data selection using sparse approximation, 2) subspace basis determination, 3) empirical subspace intersection removal and 4) minimum distance based classification. By introducing the generalized subspace model, the signals from one class are assumed to be generated from one of several subspaces. The technique is then applied to the following data sets: (a) the wooden log data for quality assessment using microwave signals; and (b) the UCI data set ARCENE. In both cases, it gives promising results when the training set is identified by a sparse representation. Moreover, by removing the empirical subspace intersection, the classification performance can be further improved.

Acknowledgment

This work was supported by the Swedish Research Council (VR) which is gratefully acknowledged.

References

[1] D. Arpit, S. Wu, P. Natarajan, R. Prasad and P. Natarajan, ridge regression based classifiers for large scale class imbalanced datasets, *Applications of Computer Vision (WACV), 2013 IEEE Workshop on* **15–17** (January 2013), 267–274.

[2] T. Balachander and R. Kothari, Kernel based subspace pattern classification, *IEEE International Joint Conference on Neural Networks* **5** (1999), 3119–3122.

[3] B.V. Dasarathy, Nearest neighbor (NN) norms: NN pattern classification techniques, *IEEE Computer Society* (December 1990).

[4] O. Besson and L.L. Scharf, CFAR matched direction detector, *IEEE Transactions on Signal Processing (TSP)* **54**(7) (2006), 2840–2844.

[5] C.M. Bishop, Pattern recognition and machine learning, Springer (2006).

[6] S. Boyd and L. Vandenberghe, Convex optimization, Cambridge University Press, Cambridge, (March 2004).

[7] H. Cevikalp, D. Larlus, M. Douze and F. Jurie, Local subspace classifiers: Linear and nonlinear approaches, *IEEE Workshop on Machine Learning for Signal Processing* (August 2007), 57–62.

[8] N. Cristianini and T.J. Shawe, An introduction to support vector machines and other kernel-based learning methods, First Edition, Cambridge University Press, 2002.

[9] E. Elhamifar and R. Vidal, Robust classification using structured sparse representation, *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on* **20–25** (June 2011), 1873–1879.

[10] A. Fhager, T. McKelvey and M. Persson, Stroke detection using a broad band microwave antenna system, *4th European Conference on Antennas and Propagation*, Barcelona, Spain, (April 2010), pp. C13P1-2.

[11] M.A.T. Figueiredo, Adaptive sparseness for supervised learning, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **25**(9) (September 2003), 1150–1159.

[12] G.H. Golub and C.F. Van Loan, Matrix computations, 3rd edition, Johns Hopkins University Press, 1996.

[13] I. Guyon, Design of experiments of the NIPS 2003 variable selection benchmark, http://www.nipsfsc.ecs.soton.ac.uk/papers/Datasets.pdf, http://www.nipsfsc.ecs.soton.ac.uk/results/?ds=arcene., 2003.

[14] T. Hastie, R. Tibshirani and J. Friedman, The elements of statistical learning: Data mining, inference, and prediction, 2nd Edition, Springer, Feb. 2009.

[15] H.K. Solanki, V.D. Prajapati and G.K. Jani, Microwave technology – a potential tool in pharmaceutical science, *International Journal of PharmTech Research* **2**(3) (July–September 2010), 1754–1761,

[16] S. Huang and J. Yang, Improved principal component regression for face recognition under illumination variations, *IEEE Signal Process Lett* **19**(4) (April 2012), 179–182.

[17] S.M. Huang and J.F. Yang, Unitary regression classification with total minimum projection error for face recognition, *IEEE Signal Processing Letters* **20**(5) (May 2013), 443–446.

[18] C.O. Kappe and A. Stadler, Microwaves in organic and medicinal chemistry, Wiley-VCH, Weinheim, 2005.

[19] S. Kraut, L.L. Scharf and L.T. McWhorter, Adaptive subspace detectors, *IEEE Transactions on Signal Processing (TSP)* **49**(1) (2001), 1–16.

[20] J. Laaksonen, Local subspace classifier, *Proceedings of ICANNs* (October 1997), 637–642.

[21] I. Naseem, R. Togneri and M. Bennamoun, Linear regression for face recognition, *IEEE Trans Pattern Analysis and Machine Intelligence* **32**(11) (July 2010), 2106–2112.

[22] I. Naseem, R. Togneri and M. Bennamoun, Robust regression for face recognition, *Pattern Recognition* **45**(1) (January 2012), 104–118.

[23] A.Y. Ng, Feature selection, L1 vs. L2 regularization, and rotational invariance, *21st ICML*, New York, USA, ACM, 2004.

[24] E. Oja, Subspace methods of pattern recognition, Research Studies Press, Letchworth and J. Wiley, 1983.

[25] M. Persson, T. McKelvey, A. Fhager, H. Lui etc., Advances in neuro diagnostic based on microwave technology, transcranial magnetic stimulation and eeg source localization, *Asia Pacific Microwave Conference* (2011).

[26] C.R. Rao, Linear statistical inference and its applications, 2nd Edition, Wiley, Dec. 2001.

[27] S.J. Raudys and A.K. Jain, Small sample size effects in statistical pattern recognition: Recommendations for practitioners, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **13**(3) (Mar 1991), 252–264.

[28] A. Rosen, M.A. Stuchly and A. Van der Vorst, Applications of RF/microwaves in medicine, *IEEE Trans on Microwave Theory and Techniques* **50**(3) (March 2002), 963–974.

[29] L.L. Scharf and B. Friedlander, Matched subspace detectors, *IEEE Transactions on Signal Processing (TSP)* **42**(8) (1994), 2146–2157.

[30] B. Scholkopf and A.J. Smola, Learning with kernels: Support vector machines, regularization, optimization, and beyond, First Edition, MIT Press, December, 2001.

[31] R. Tibshirani, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* **58** (1996), 267–288.

[32] V. Vapnik, The nature of statistical learning theory, 2nd Edition, Wiley, 1999.

[33] P. Vincent and Y. Bengio, K-local hyperplane and convex distance nearest neighbour algorithms, *Proceedings of NIPS* (2001), 985–992.

[34] T.C. Williams, J.M. Sill and E.C. Fear, Breast surface estimation for radar-based breast imaging systems, *IEEE Transactions on Biomedical Engineering* **55**(6) (May 2008), 1678–1686.

[35] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry and Y. Ma, Robust face recognition via sparse representation, *IEEE Transaction on Pattern Analysis and Machine Intelligence* **31**(2) (February 2009).

[36] S. Yan, J. Liu, X. Tang and T.S. Huang, A parameter-free framework for general supervised subspace learning, *Information Forensics and Security* **2**(1) (2007), 69–76.

[37] Y. Yu and T. McKelvey, A subspace learning algorithm for microwave scattering signal classification with application to wood quality assessment, *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on* (September 2012), 23–26.

[38] O. Zeitouni, J. Ziv and N. Merhav, When is the generalized likelihood ratio test optimal? *IEEE Transactions on Information Theory (TIT)* **38**(5) (1992), 1597–1602.

[39] D. Zou, Local subspace classifier in reproducing kernel hilbert space, *Advances in Multimodal Interfaces – ICMI, Lecture Notes in Computer Science* **1948** (2000), 434–441.

[40] B. Xua, P. Guo and C.L.P. Chen, An adaptive regularization method for sparse representation, *Integrated Computer-Aided Engineering* **21** (2004), 91100.